Depth-Aware Concealed Crop Detection in Dense Agricultural Scenes -Supplementary Material-

Liqiong Wang[1,2†], Jinyu Yang[3,4†], Yanfu Zhang[5], Fangyi Wang[1,2∗], Feng Zheng[3∗]
[1]Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University
[2]College of Computer and Information Technology, China Three Gorges University
[3]Southern University of Science and Technology [4]University of Birmingham
[5]Departmental of Computer Science, College of William and Mary
{liqiong.wang11,jinyu.yang96}@outlook.com yzhang105@wm.edu fy_wang@ctgu.edu.cn f.zheng@ieee.org

1. More Details of CCD

The motivation of CCD lies in the challenges posed by dense plantings in agricultural scenes. Close plant growth makes it difficult to distinguish individual crops or detect concealed objects, leading to inaccurate crop counting, yield estimation, and so on. This can significantly impact agricultural management and decision-making. Thus, we aim to address these challenges by leveraging depth cues to enhance the detection of concealed objects and individual crops. By incorporating depth awareness into the detection process, the method can better handle occlusions and dense plantings, allowing for more accurate and robust crop detection even in challenging scenarios. The proposed CCD is definitely useful and has potential in broad agriculture-related applications, *e.g.*, fine weeding and pruning, yield estimation, and automatic agricultural robot systems.

2. More Dataset Statistics

In this section, we report the ratio of objects with different sizes, shown in Tab. 1. Small object (SO) typically refers to those that occupy a relatively small area ($\leq 1\%$) within an image, often leading to challenges in detection due to their limited visual information. Dense object (DO) refers to those closely packed together, making it difficult to separate and identify individual instances.

3. More Ablation Study

In this section, we perform comprehensive ablation experiments on various modules to extensively validate the effectiveness of the three main modules in RISNet, *i.e.*, CFE, DFD, and IFR, as well as the rationality behind the design of each module.

† Equal contribution. ∗ Corresponding author.
This work was done during Liqiong Wang visited SUSTech VIP lab.

Target Size	R$\leq 0.2\%$	0.2% <R$\leq 1\%$	1% <R
Ratio	75.5%	21%	3.5%

Table 1. Statistics on the ratio of objects with different sizes.

Method	$S_\alpha \uparrow$	$F^\omega_\beta \uparrow$	$E_\theta \uparrow$
(a) ResNet-50+Conv	0.842	0.762	0.961
(b) ResNet-50+ASPP	0.843	0.763	0.950
(c) Res2Net-50+Conv	0.855	0.785	0.964
(d) Res2Net-50+ASPP	0.852	0.781	0.957
(e) PVT+Conv	0.863	0.793	**0.967**
(f) RISNet	**0.866**	**0.803**	**0.967**

Table 2. Ablation study of CFE Module.

Method	$S_\alpha \uparrow$	$F^\omega_\beta \uparrow$	$E_\theta \uparrow$
(a) MFF→Concat+w/o RFD	0.861	0.790	0.965
(b) w/o MFF	0.864	0.795	**0.967**
(c) w/o RFD	0.863	0.799	0.966
(d) MFF→Concat	0.863	0.795	0.966
(e) RISNet	**0.866**	**0.803**	**0.967**

Table 3. Ablation study of DFD Module.

3.1. Effect of CFE Module

We provide detailed information on the ablation experiments of the CFE module in Tab. 2. Our CFE module is mainly composed of two parts, *i.e.*, the PVT-based encoder and the ASPP module. These modules are designed to capture the feature information of densely packed objects in

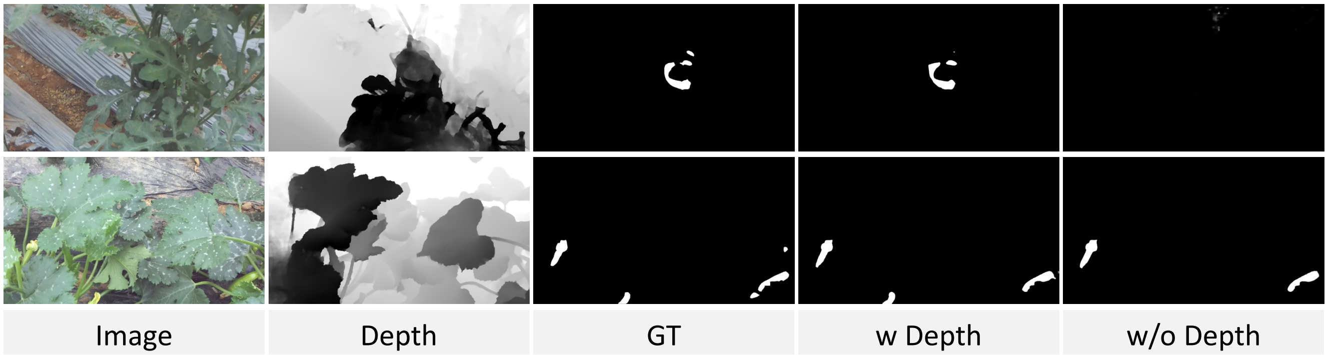


Figure 1. Visual comparison of our RISNet with/without depth.

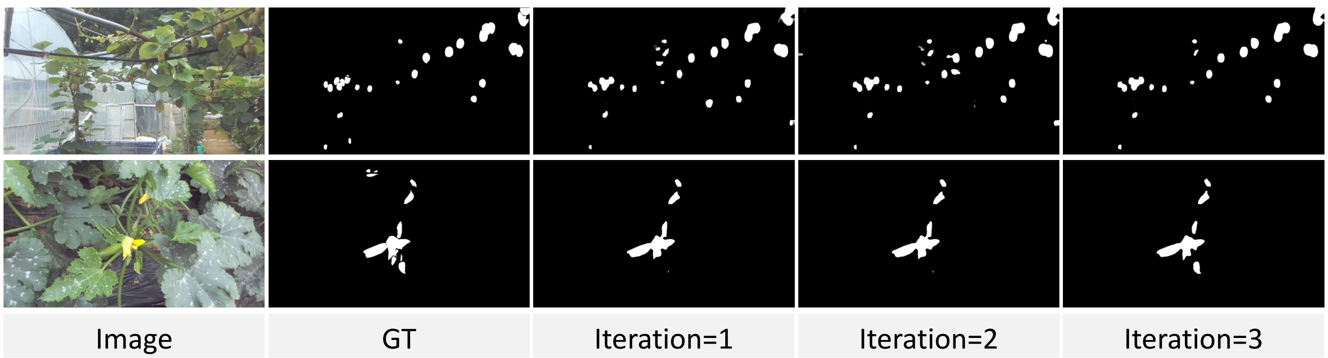


Figure 2. Visual comparison of different iterations in our IFR module.

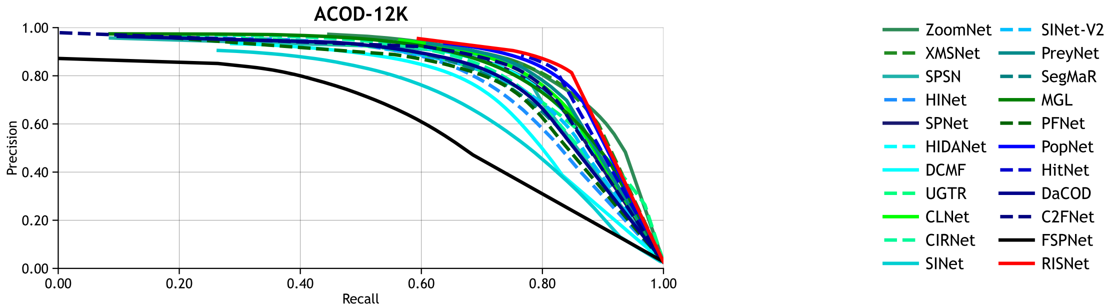
Method	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$E_\theta \uparrow$
(a) itration 1+w/o FAF	0.850	0.785	0.949
(b) itration 1	0.858	0.792	0.961
(c) w/o GGA	0.866	0.795	**0.967**
(d) RISNet	**0.866**	**0.803**	**0.967**

Table 4. Ablation study of IFR Module.

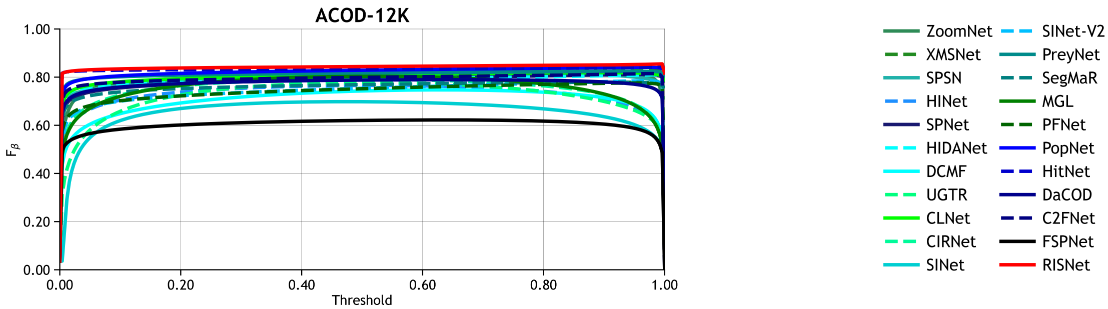
CCD. To verify the effectiveness of each module, we perform ablation experiments on them separately. In (b) and (d), we replace the encoder backbone from PVT to ResNet-50 [8] and Res2Net-50 [7], respectively. Building upon this, (a) and (c) further substitute the ASPP module with a simple convolution operation. In (e), we replace the ASPP module with a convolution while keeping the backbone unchanged. The comparison between (f) and (a), (b), (c), (d) indicates that replacing the backbone leads to varying degrees of performance decline, demonstrating that PVT, as the backbone, is more suitable for our task. The comparison between (e) and (f) reveals a decrease in F_β^ω after replacing the ASPP module, indicating a reduction in the model's prediction accuracy. This is because, in contrast to convolution, ASPP can capture object information through features with different scales, thereby aiding the model in achieving more accurate detection results.

3.2. Effect of DFD Module

Tab. 3 presents detailed information on the ablation experiments of the DFD module. Our DFD module consists of the MFF module and the RFD module, and we sequentially conduct ablation experiments to validate the effectiveness of each module. In (b) and (c), we remove the MFF module and the RFD module, respectively. In (d), a simple concatenation is used to fuse information from the two modalities. (a) is obtained by removing the RFD module from (d). (b) essentially uses only single-modal information. Comparing (e) with (b), the introduction of depth information significantly improves F_β^ω, indicating that depth information contributes to better object localization in complex environments. It is worth noting that comparing (e) with (d) and (a) with (c), we find that our MFF module can better integrate information from the two modalities than concatenation. Even comparing (b) with (d), when concatenating multi-modal information, the metrics relative to using only single-modal information slightly decrease. This is because, in the case of concatenation, the model tends to rely more on RGB modality information. The decrease in metrics for (c) compared to (e) also indicates the effective-



(a) PR curves.



(b) F_β curves.

Figure 3. PR and F_β curves of the proposed RISNet and recent SOTA algorithms on CCD.

Model	Publications	NLPR				NJU2K				STERE				SIP			
		$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_m\uparrow$
CoNet[9]	ECCV20	0.027	0.903	0.911	0.943	0.046	0.902	0.896	0.926	0.037	0.909	0.905	0.941	0.058	0.887	0.860	0.911
DASNet[23]	MM20	**0.021**	**0.929**	0.929	0.960	0.042	0.911	0.902	0.935	0.037	0.915	0.910	0.939	0.051	0.900	0.877	0.918
RD3D[1]	AAAI21	0.022	0.927	**0.930**	0.959	0.036	0.923	0.916	0.941	0.037	0.917	0.911	0.939	0.048	0.906	0.885	0.918
JLDCF[6]	TPAMI21	0.022	0.925	0.925	0.955	0.041	0.912	0.902	0.936	0.040	0.913	0.903	0.934	0.049	0.903	0.880	0.918
BIANet[22]	TIP21	0.023	0.924	0.926	0.956	0.036	0.929	0.917	0.942	0.039	0.912	0.905	0.935	**0.047**	0.904	0.887	0.920
BBSNet[20]	TIP21	0.023	0.927	**0.930**	0.953	0.035	0.931	0.920	0.941	0.041	0.919	0.908	0.931	0.055	0.902	0.879	0.910
DSNet[18]	TIP21	0.024	0.925	0.926	0.951	0.034	0.929	0.921	0.946	0.036	**0.922**	**0.914**	0.941	0.052	0.899	0.876	0.910
UTANet[24]	TIP21	**0.020**	0.928	**0.932**	**0.964**	0.037	0.915	0.902	0.945	**0.033**	**0.921**	0.910	**0.948**	0.048	0.897	0.873	0.925
DCF[10]	CVPR21	0.022	0.918	0.924	0.958	0.036	0.922	0.912	0.946	0.039	0.911	0.902	0.940	0.052	0.899	0.876	0.916
DSA2F[17]	CVPR21	0.024	0.897	0.918	0.950	0.039	0.901	0.903	0.923	0.036	0.898	0.904	0.933	-	-	-	-
SPNet[26]	ICCV21	**0.021**	0.925	0.927	0.959	**0.028**	**0.935**	**0.925**	**0.954**	0.037	0.915	0.907	0.944	**0.043**	**0.916**	**0.894**	**0.930**
TriTrans[13]	MM21	**0.020**	0.923	0.928	0.960	0.030	0.926	0.920	0.925	**0.033**	0.911	0.908	0.927	**0.043**	0.898	0.886	0.924
C2DFNet[21]	TMM22	**0.021**	0.926	0.928	0.956	-	-	-	-	0.038	0.911	0.902	0.938	0.053	0.894	0.782	0.911
MVSalNet[25]	ECCV22	0.022	**0.931**	**0.930**	0.960	0.036	0.923	0.912	0.944	0.036	**0.921**	**0.913**	0.944	-	-	-	-
SPSN[12]	ECCV22	0.023	0.917	0.923	0.956	0.032	0.927	0.918	**0.949**	**0.035**	0.909	0.906	0.941	**0.043**	0.910	0.891	**0.932**
HiDAnet[19]	TIP23	**0.021**	**0.929**	**0.930**	**0.961**	**0.029**	**0.939**	**0.926**	**0.954**	**0.035**	**0.921**	0.911	**0.946**	**0.043**	**0.919**	**0.892**	0.927
Ours		**0.016**	**0.939**	**0.937**	**0.971**	**0.027**	**0.941**	**0.928**	**0.955**	**0.031**	**0.924**	**0.917**	**0.949**	**0.038**	**0.924**	**0.900**	**0.936**

Table 5. Detailed comparison results of different methods on RGB-D SOD task. The best three results are highlighted in **red**, **blue** and **green**.

ness of the RFD module.

3.3. Effect of IFR Module

The details of the ablation experiments for the IFR module are shown in Tab. 4. Unlike the previous two modules, the IFR module is composed of iterative optimization and final low-level feature fusion optimization. In (b), we eliminate iterative optimization. In (c), based on (b), we remove the fusion of low-level features and directly output the prediction result. In (d), we remove the GGA module during the iterative optimization process, meaning that the coarse prediction map from the previous stage is no longer used to assist in locating objects in the next stage. Comparing (d) with (a), (b), and (c), the significant decrease in metrics indicates that our iterative optimization strategy is very beneficial for the model. The comparison between (a) and (b) demonstrates the effectiveness of our FAF module, low-level features contain more geometric information, and fusing this information helps optimize our prediction results. The comparison between (d) and (c) highlights the importance of GGA. With the assistance of GGA, our model can better locate the position of small objects, allowing the model to focus on the region of interest and aiding in the detection of challenging objects.

4. More Comparisons

4.1. Effectiveness of Depth

In Fig. 1, we illustrate the utility of depth information in aiding RISNet in object detection.

4.2. Visual comparison of each iteration

We show the output of each iteration in Fig. 2.

4.3. PR & F_β curves on CCD

In Fig. 3, we show the PR & F_β curves of different methods on CCD. The red curve represents our method.

5. Experiments on SOD

5.1. Datasets

For the RGB-D SOD task, we follow established practices by [10, 20, 26], selecting 1485 samples from NJU2K [11] and 700 samples from NLPR [16], totaling 2185 samples for training. Subsequently, we assess the performance of our model on widely used datasets, including NLPR [16], NJU2K [11], STERE [15], and SIP [4].

5.2. Evaluation Metrics

Following [26], we use the widely adopted metrics mean absolute error M, max F-measure F_β [14], max E-measure E_m [3] [5], and structure measure S_α [2] as our evaluation criteria.

5.3. Comparisons with State-of-the-arts

We compare our proposed RISNet with several existing RGB-D SOD methods. As depicted in Tab. 5, our model still achieves superior results. This further demonstrates the generalization capability of our model and underscores the superiority of our framework.

References

[1] Qian Chen, Ze Liu, Yi Zhang, Keren Fu, Qijun Zhao, and Hongwei Du. Rgb-d salient object detection via 3d convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1063–1071, 2021. 3

[2] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017. 4

[3] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018. 4

[4] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems*, 32 (5):2075–2089, 2020. 4

[5] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *Scientia Sinica Informationis*, 6(6), 2021. 4

[6] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, Qijun Zhao, Jianbing Shen, and Ce Zhu. Siamese network for rgb-d salient object detection and beyond. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5541–5559, 2021. 3

[7] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[9] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate rgb-d salient object detection via collaborative learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 52–69. Springer, 2020. 3

[10] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated rgb-d salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9471–9481, 2021. 3, 4

[11] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *2014 IEEE international conference on image processing (ICIP)*, pages 1115–1119. IEEE, 2014. 4

[12] Minhyeok Lee, Chaewon Park, Suhwan Cho, and Sangyoun Lee. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In *European Conference on Computer Vision*, pages 630–647. Springer, 2022. 3

[13] Zhengyi Liu, Yuan Wang, Zhengzheng Tu, Yun Xiao, and Bin Tang. Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4481–4490, 2021. 3

[14] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255, 2014. 4

[15] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 454–461. IEEE, 2012. 4

[16] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 92–109. Springer, 2014. 4

[17] Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi Li. Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1407–1417, 2021. 3

[18] Hongfa Wen, Chenggang Yan, Xiaofei Zhou, Runmin Cong, Yaoqi Sun, Bolun Zheng, Jiyong Zhang, Yongjun Bao, and Guiguang Ding. Dynamic selective network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:9179–9192, 2021. 3

[19] Zongwei Wu, Guillaume Allibert, Fabrice Meriaudeau, Chao Ma, and Cédric Demonceaux. Hidanet: Rgb-d salient object detection via hierarchical depth awareness. *IEEE Transactions on Image Processing*, 32:2160–2173, 2023. 3

[20] Yingjie Zhai, Deng-Ping Fan, Jufeng Yang, Ali Borji, Ling Shao, Junwei Han, and Liang Wang. Bifurcated backbone strategy for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:8727–8742, 2021. 3, 4

[21] Miao Zhang, Shunyu Yao, Beiqi Hu, Yongri Piao, and Wei Ji. C2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection. *IEEE Transactions on Multimedia*, 2022. 3

[22] Zhao Zhang, Zheng Lin, Jun Xu, Wen-Da Jin, Shao-Ping Lu, and Deng-Ping Fan. Bilateral attention network for rgb-d salient object detection. *IEEE transactions on image processing*, 30:1949–1961, 2021. 3

[23] Jiawei Zhao, Yifan Zhao, Jia Li, and Xiaowu Chen. Is depth really necessary for salient object detection? In *Proceedings of the 28th ACM international conference on multimedia*, pages 1745–1754, 2020. 3

[24] Yifan Zhao, Jiawei Zhao, Jia Li, and Xiaowu Chen. Rgb-d salient object detection with ubiquitous target awareness. *IEEE Transactions on Image Processing*, 30:7717–7731, 2021. 3

[25] Jiayuan Zhou, Lijun Wang, Huchuan Lu, Kaining Huang, Xinchu Shi, and Bocong Liu. Mvsalnet: Multi-view augmentation for rgb-d salient object detection. In *European Conference on Computer Vision*, pages 270–287. Springer, 2022. 3

[26] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4681–4691, 2021. 3, 4